Video: Boxplots Intro (4:40)

(0:00):
Let's talk about boxplots.

(0:04):
Boxplots are unique in that they were introduced fairly recently. They were introduced by a mathematician, John Tuckey in 1969. The boxplots we will use in class are called Tukey box plots. Another nomenclature for this graph is a 'box and whisker' plot, and when you see one, you'll understand how it got its name. It can be either vertical or horizontal.

(0:30):
A box plot presents 5 values from your data. It is built off the median, and so everything is based on that. To build one of these, sort your data into quartiles (quarters), and you plot the max, the 75% value, the median, the 25% value, and the minimum value. IQR is the interquartile range and that's the middle 50% and you get that by subtracting the 75% value from the 25% value.

(1:02):
This is a boxplot that was built using the wakeup time from the sleep diary data. Just like a line chart, the x axis is the independent variable (the student section number) and the y axis is the dependent variable, just like in a line or scatter plot. You can see where the nomenclature (boxplot) came from, but what about the whiskers? Turn it horizontal and the black lines on these probably look like whiskers…

(1:33):
Let's break it down into its component pieces. The graph is built on the median, which is the middle line in the box, and is red in MATLAB. The top of the box is the 75% data value. The bottom of the box is the 25% value. The top of the whisker is the maximum value, if it is within 1.5 * IQR, and the minimum value is the bottom of the whisker with the same limitation. Whiskers are black in MATLAB. But what about those red crosses? Any values larger or smaller than 1.5 * IQR is called an outlier. They are real values, just outside the whisker. This limit (1.5 * IQR) makes this a Tukey boxplot, while other versions don't have that limit. With a boxplot, you have no idea how many data points you have.

(2:30):
To analyze these, notice the differences in sizes in the upper and lower boxes – section 0's box is much smaller on the bottom, indicated that the data points between 25% and 50% are fairly tightly grouped. But that size is about the same as the lower and upper whisker, so those 3 spreads seem to be evenly distributed, and the 50% to 70% is all spread out. Section 3 has equal sized upper and lower box sizes, and fairly same sized whiskers, but has numerous outliers. Does this data make sense? This is wakeup time, with Section 0 as the instructors and the rest are students. The bottom of the box is early for Section 0 (before 6am) and seems appropriate (between 6:30 and 7:45) and the 75% value also seems appropriate (between 9 and 10 am). 50% of the values are between the bottom and top of the box.

(3:55):
To get even more information from the graph, we add a 'notch'. Notice that the notch is connected to the median. The software calculates the middle 5%, which is indicated by the width of the notch. When comparing 2 data sets, like the median of Section 0 and Section 1, there is no overlap of the notch, so there might be evidence of statistical significance.

(4:25):

In summary, you can get a lot of information from a boxplot, but it focuses on the maximum, the 75% value, the median, the 25% value and the minimum value and how they are spread out.